High Dimensional Mode Hunting Using Pettiest Components Analysis

Tianhao Liu, Daniel Andrés Díaz-Pachón, J. Sunil Rao, and Jean-Eudes Dazard,

Abstract—Principal components analysis has been used to reduce the dimensionality of datasets for a long time. In this paper, we will demonstrate that in mode detection the components of smallest variance, the pettiest components, are more important. We prove that when the data follows a multivariate normal or Laplace distribution, we obtain boxes of optimal volume by implementing "pettiest component analysis", in the sense that their volume is minimal over all possible boxes with the same number of dimensions and fixed probability. This reduction in volume produces an information gain that is measured using active information. We illustrate our results with a simulation and a search for modal patterns of digitized images of hand-written numbers using the famous MNIST database; in both cases pettiest components work better than their competitors. In fact, we show that modes obtained with pettiest components generate better written digits for MNIST than principal components.

Index Terms—Active information, Bump hunting, Dimension reduction, Mode hunting, Principal components analysis.

1 Introduction

Principal components analysis is a widely used learning tool, particularly in unsupervised learning, to reduce the dimension of the space by projecting on the orthogonal rotation that maximizes the variance, especially in $p \gg n$ datasets [1, Ch. 8]. In principal components regression, most practitioners prefer to discard the components of the input with the smallest variance, using just the first few principal components. Discarding the components of smallest variance does give satisfying results in some situations. In fact, Artemiu and Li gave conditions under which the first few leading principal components have a higher probability of correlating with the dependent variable than the small-variance ones [2]. However, even though using principal components regression is not totally invalid, some caution is needed when it is implemented. As Jolliffe demonstrated in a classic paper, we can easily find some ordinary examples in which small-variance components become equally important, if not more, than large-variance ones [3]. Hadi and Ling have also presented three cautionary notes on using principal components regression, mainly due to issues arising from multicollinearity [4]. In the end, the strongest reason to question principal components regression is that the dependent variables are never used in principal components analysis. In the words of Cox, it is hard to see any reason "why the dependent variable should not be closely tied to the least important principal components" [5, p. 272]. Here we will call those least important principal components *pettiest components*.

JSR was partially supported by NSF grant DMS-1915976 and NIH grants U54 MD010722 and UL1 TR000460. DADP was partially supported by grant AWD-005895 from the Walter Bradley Center for Natural and Artificial Intelligence. (Corresponding author: Daniel Andrés Díaz-Pachón.)

- *T. Liu, D. A. Díaz-Pachón and J. S. Rao are with the Division of Biostatistics, University of Miami, Miami, Fl, 33136*
  E-mail: txl646@miami.edu, Ddiaz3@miami.edu, JRao@miami.edu
- *J-E. Dazard was with the Center of Proteomics and Bioinformatics at Case Western Reserve University, Cleveland, OH, 44106*
  E-mail: jean-eudes.dazard@case.edu

Although these arguments are informative, to the best of our knowledge, few have tried to make a systematic use of the pettiest components in their own right. That is, most criticisms have focused on the negative aspect of finding isolated counter-examples showing that principal components regression is not desirable, based on the underlying assumption that the leading principal components are better, but the positive aspect of systematically selecting the pettiest components has received little attention. The exception seems to be a recent study by Sando and Hino where the authors propose a modal principal component analysis whose goal is to develop a robust PCA method. Through a kernel function they estimate the mode and use this estimation in order to find the pettiest components, that they call them "minor components" [6].

Our reasoning goes the other way around —we take advantage from these pettiest components in order to find modes. In fact, a modification of a mode hunting algorithm called fast-PRIM (the topic of Section 3 in this article) will show that the smallest regions of the same probability are achieved using the pettiest components. Intuitively, the result presented here follows this very simple idea: In a space $S \subset \mathbb{R}^p$, define a β**-mode** of a continuous distribution as the region of the space with the smallest volume constrained to have a probability β. Assume the underlying distribution of a p-dimensional vector $\mathbf{x}$ is a multivariate normal, and that we project it to a p'-dimensional space, with $p' < p$. Then the box that minimizes the size, among all possible boxes of probability β, is the one projected over the subspace of the orthogonal variables with the smallest variance.

This work can be seen as a continuation of [7], where the optimality, in terms of smallest volume, of the p-dimensional box centered around the mean in the direction of the eigenvectors was proved for a multivariate normal distribution. The next obvious step, the reduction of dimensionality to p' dimensions, is accomplished here; but contrary to common practice, we show that the optimal p'-box is obtained when the p-dimensional box in the previous step

is projected to the subspace of the p' pettiest components.

Several questions and comments arise from this situation. First, finding a general solution for the set of minimum volume in the class $\mathcal{C}$ of all the Borel sets with probability β in $\mathbb{R}^d$ seems in general intractable; therefore we consider the next big thing: the more manageable class $\mathcal{C}' \subset \mathcal{C}$ of all hyper-rectangles $I_1 \times \cdots I_p$, where I_i are intervals in $\mathbb{R}$. Second, we define *a* β-mode and not *the* β-mode, because there can be more than one Borel set with identical hypervolume and probability β in the space; in fact, in many situations not only global but local β-modes are of interest. Third, we talk about β-modes in general, instead of simply modes, because we are interested on regions of positive probability. And fourth, out of the previous considerations, there are no spaces without β-modes; i.e., even if S is of finite volume and the underlying distribution is uniform we have uncountable β-modes, though these will not be very informative; on the other hand, if S has infinite Lebesgue measure, a continuous distribution in S will have at least a β-mode.

Why the interest in the β-regions with the smallest volume? An answer among many approaches lies in the fact that such regions might contain a large amount of relative Shannon information among all the regions with probability β (see Subsection 2.2).

In spite of the p-dimensional support of a continuous distribution having uncountably infinite regions of probability β, even when $p = 1$, the smallest region of probability β will tell us that the data is highly concentrated inside such region. And there is a vast number of applications where this is of interest. For instance, in medical image recognition a high concentration of points around a small region can help on tumor detection; or in more accurate predictions based on closeness to a given target, as in the Netflix recommendations algorithm. Many more come easily to mind.

Yet, mode detection continues being a difficult problem, specially in multivariate settings. One of the most famous ways to approach it, specially in computer vision, is via the mean-shift algorithm by kernel density estimation [8]. It works well when $p = 2$ having amenable asymptotic properties [9], [10]. However, it becomes very slow when $p > 2$, though the speed improves when the shape of the density is known [11]. More recently, Ruzankin and Logashov introduced a fast mode estimator with time complexity $O(pn)$, whereas other estimators time complexity is $O(pn^2)$, where n is the number of observations [12].

2 Basic Notions

2.1 Components

Suppose there is a p-dimensional dataset of n observations $\{x_1, \ldots, x_p\}_1^n$ with covariance matrix Σ. The problem is to find a vector $\mathbf{a}$ maximizing $\mathbf{a}^T\Sigma\mathbf{a}$, subject to $\mathbf{a}^T\mathbf{a} = 1$. By Lagrange multipliers, the original optimization problem is equivalent to solve the eigenvalue question $\Sigma\mathbf{a} = \lambda\mathbf{a}$. From the solution we get the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$, and respective eigenvalues $\lambda_1, \ldots, \lambda_k$. The original p-dimensional input space will be denoted by $\mathcal{X}(p)$, and the space after the rotation in the direction of its p eigenvectors, will be denoted by $\mathcal{X}'(p)$.

We will call these eigenvectors *components*, and their corresponding eigenvalue will be their variance. Then, sorting the components in order of increasing variance, the k *principal components* will be the k components with the largest variances; and the k *pettiest components* will be the k components with the smallest variance.

2.2 Optimality of the β-mode

We define first a continuous version of unimodality and then define active information in order to prove the optimality of the β-mode of a unimodal distribution ϕ in terms of Shannon information.

Definition 1 (β-unimodality). We say that a continuous distribution ϕ is β-unimodal with β-mode B if for any other β-mode B' the volume of the symmetric difference $\mathrm{Vol}(B \triangle B')$ is 0. In this case either B or B' can be taken as the β-mode.

Definition 2 (Active information). Let ϕ and ψ be two continuous distributions on $\mathcal{X}(p)$. The active information of ϕ with respect to ψ for an event $R \in \mathcal{X}(p)$ is defined as

$$\mathbf{I}_+(R) := \log \frac{\phi(R)}{\psi(R)}, \tag{1}$$

provided R has positive probability under ϕ and ψ.

Thus, active information represents the information change induced by ϕ with respect to the baseline distribution ψ [13]. Active information was introduced in the context of search problems and has also been used to detect modes in multivariate analyses [16], [17]. We present it here in order to prove the optimality of the β-mode in terms of Shannon information.

Theorem 1 (Optimality of the β-mode). Let $\mathcal{X}(p)$ be bounded. Without loss of generality, assume $\mathcal{X}(p) = [0,1]^p$. Let ϕ be a continuous distribution on $[0,1]^p$. Then $B \subset [0,1]^p$ is the unique β-mode of ϕ if and only if B maximizes the active information of ϕ with respect to a uniform distribution $\mathbf{U}$ among all the events of $[0,1]^p$ with probability β.

Proof: Take B and B' two events such that $\phi(B) = \phi(B') = \beta$, with respective volumes v and v'. Then the active information of ϕ with respect to $\mathbf{U}$ for the event B is

$$\mathbf{I}_+(B) := \log \frac{\phi(B)}{\mathbf{U}(B)} = \log \frac{\beta}{v}, \tag{2}$$

and the active information of ϕ with respecto to $\mathbf{U}$ for the event B' is

$$\mathbf{I}_+(B') := \log \frac{\phi(B')}{\mathbf{U}(B')} = \log \frac{\beta}{v'}. \tag{3}$$

Therefore

$$\begin{aligned} \mathbf{I}_+(B) - \mathbf{I}_+(B^*) &= \log \frac{\beta}{v} - \log \frac{\beta}{v'} \\ &= \log \frac{v'}{v}, \end{aligned} \tag{4}$$

which is positive if and only if $v' > v$. □

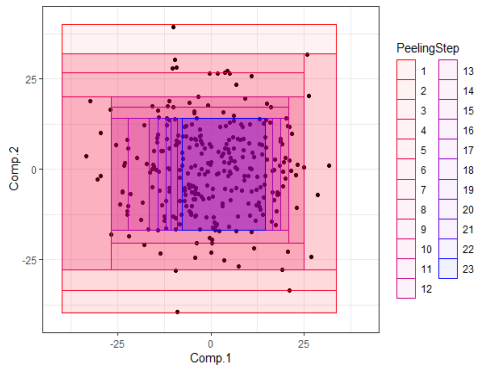


Figure 1: The peeling procedure in PRIM.

2.3 Patient rule induction method

The *patient rule induction method* (PRIM) is a greedy algorithm designed for bump hunting [18]. A bump can be roughly defined as a region on the response variable with higher mean, compared to other places.

Assume we have a dataset $\{y, x\}_1^n$ where $x = \{x_1, x_2, \ldots, x_p\}$ is a continuous random vector. $f(x)$ is the target function. The input domain $\mathcal{X}(p)$ is here called S. Defining $\bar{f}_B$ and $\bar{f}$ as

$$\bar{f}_B = \frac{\int_B f(x)p(x)dx}{\int_B p(x)dx},$$
$$\bar{f} = \int_S f(x)p(x)dx,$$

where $p(x)$ is the density function of x, the method aims to find a box $B \subset S$ in which $\bar{f}_B/\bar{f}$ is maximized under the constraint that B is not too small. In fact, the probability of B is defined at the outset by a tuning meta-parameter β for the smallest permitted box size. In practice, $\bar{f}_B/\bar{f}$ will be replaced by its corresponding estimator.

The algorithm is divided in three stages. First there is a *peeling*: Beginning with the whole input space S, a whole class of eligible boxes for removal is set up. This class is made of sets of the form

$$b_{j-} = \{x \mid x_j < x_{j(\alpha)}\},$$
$$b_{j+} = \{x \mid x_j > x_{j(1-\alpha)}\},$$

where $x_{j(\alpha)}$ is an α-quantile. If a box b^* is in the eligible set, and without it the remaining $S \setminus b^*$ will give the maximum average value of response variable, b^* is the specific box selected to remove in this loop. Formally we write

$$b^* = \arg\max_b \text{ave}[y_i \mid x_i \in B - b],$$

where ave is the average. Update the new box as $B = S \setminus b^*$ and run the same procedure on B. The process is iterated until the final box B has probability β. The peeling procedure is shown in Fig. 1.

After peeling, a second step called *pasting* is performed in order to correct for the greediness of peeling. We will not consider pasting here. Finally, the third step is called *covering*, which means that after the peeling stage, the final box B is deleted from S, and then the whole procedure of peeling is repeated in $S \setminus B$.

PRIM has at least three shortcomings. First, it is computationally expensive; second, it does not behave well in the presence of collinearity; and third, even in a small number of dimensions, it cannot detect multiple bumps [19]. For this reason, [20] developed a modified algorithm called *local sparse bump hunting* (LSBH) in which the patient rule induction method is subsumed. The LSBH algorithm uses a recursive partition algorithm, say classification and regression trees, in order to divide the space $\mathcal{X}(p)$ in several regions $S_1, \ldots, S_R$; then applies sparse principal components analysis inside each particular region S_i in order to reduce the dimension of each region in the partition; and finally applies the patient rule induction method to each rotated and projected subspace induced by S_i. [21] applied LSBH to detect heterogeneity of colon tumors, dividing colon cancer patients into two subpopulations with different genetic/molecular profiles in all stages of cancer.

2.4 FastPRIM

Elaborating on LSBH, a new modified algorithm called *fastPRIM* was developed in order to find the minimal volume of boxes with probability β when the distribution is a multivariate normal [7, Algorithm 4]. More accurately, the modes can be defined as β-modes; i.e., contiguous regions, not necessarily unique, with the smallest volume and probability β. Since the mode and the mean of the normal distribution coincide, mode hunting in this case is equivalent to finding a box of probability β centered around the mean. In fastPRIM the space is first rotated in the direction of its eigenvectors, and then the patient rule induction is applied in the direction of the rotation. It turns out that in this setting the whole peeling and pasting iteration is reduced to one single step. That is, fastPRIM chooses the centralized box such that each side of the box is parallel to an axis of the input space and whose vertices are located at the quantiles $2^{-1}(1 \pm \beta_T^{1/p})$ of the corresponding variable, where $\beta_T = \sum_{k=1}^t \beta(1-\beta)^{k-1} = 1 - (1-\beta)^t$ is the probability measure after t steps of covering. As a result, we obtain a rectangular Lebesgue set, or say a square set in probability, centralized at the zero point.

Therefore, from a sampling viewpoint fastPRIM is consistent: calculating the average of all the points inside the final centered box and taking this average as the center of our box, we know by the law of large numbers that the final box will be centered around the origin. Since the mean and the mode coincide in the normal distribution, as $n \to \infty$, the procedure is approaching a box whose center is the true mode. This is so with or without dimension reduction.

3 Pettiest components analysis with fastPRIM

[7, Proposition 1] proved that fastPRIM satisfies the following optimality property: when the data is multivariate normal, say $\mathcal{N}(\mathbf{0}, \Sigma)$, the box with the smallest volume subject to have probability β is found in the direction of the rotation of the p principal components; that is, the box centered around the mean of probability β in $\mathcal{X}'(p)$. However, they did not prove any result dealing with dimension reduction. This article goes one step further: it shows that if we are

going to consider reduction of dimensionality, say to $p' < p$, we should consider pettiest components instead of principal components.

Starting thus with the space $\mathcal{X}(p)$, let B_i be the final box obtained by fastPRIM on an input space $\mathcal{X}_i'(p')$, where i indicates a specific way of choosing p' variables from p variables. The collection of all such boxes will be $\mathcal{B}$. The Lebesgue measure of the box $B \in \mathcal{B}$ with probability measure β is $\text{Vol}(B|\beta)$. In the input space $\mathcal{X}_i'(p')$ spanned by the pettiest components, we write that specific box as $\mathbf{B}$.

Theorem 2. Let X be a p-random vector such that its components $X_1, \ldots, X_p$ are independent of each other and have normal distribution $\mathcal{N}(0, \sigma_1^2), \ldots, \mathcal{N}(0, \sigma_p^2)$, respectively. Apply fastPRIM on each possible projection space $\mathcal{X}_i'(p')$ with same probability β. Then,

$$\arg\min_{B \in \mathcal{B}} \text{Vol}(B|\beta) = \mathbf{B}.$$

Proof: . There are $\binom{p}{p'}$ ways to choose p' dimensions from p. For $X_1, \ldots, X_p$ following independent normal distributions, we write $X_j \sim \mathcal{N}_j(0, \sigma_j^2)$, $(j = 1, \ldots, p)$. From the fact that the final box B in fastPRIM is a square in probability, the marginal probability measure of every $B \in \mathcal{B}$ is $\beta^{1/p'}$. Because all boxes $B \in \mathcal{B}$ are centralized at zero, the Lebesgue measure of X_j, subject to a probability $\beta^{1/p'}$, can be calculated by the equation $P\{-k\sigma_j < X_j < k\sigma_j\} = \beta^{1/p'}$. Here k only depends on $\beta^{1/p'}$ and it has nothing to do with j. Therefore, the edge length of B_i in the X_j direction is $2k\sigma_j$. Then,

$$\text{Vol}(B_i|\beta) = \prod_{j^{(i)}=1}^{p'} 2k\sigma_{j^{(i)}} = (2k)^{p'} \prod_{j^{(i)}=1}^{p'} \sigma_{j^{(i)}},$$

where $j^{(i)}$ denotes the p' variables in the choice of B_i. Since a fix β implies a fix k, the minimum of $\text{Vol}(B_i|\beta)$ can thus be obtained by minimizing $\prod_{j^{(i)}=1}^{p'} \sigma_{j^{(i)}}$. But this only requires the $\sigma_{1^{(i)}}, \ldots, \sigma_{p'^{(i)}}$ to have the smallest values, which is achieved by choosing the $X_{1^{(i)}}, \ldots, X_{p'^{(i)}}$ with the smallest variances. That is, the p' pettiest components. The corresponding box found by fastPRIM with pettiest components is $\mathbf{B}$. □

The basic idea behind the proof is to reduce the box to the marginals with same probability measure and then to prove a smaller variance corresponding to a smaller Lebesgue measure. Using Theorem 2, we can easily show that the same conclusion is true for any multivariate normal distribution input under fastPRIM.

Theorem 3. Let X be a p-random vector distributed as $\mathcal{N}_p(0, \Sigma)$. Apply fastPRIM on each possible projection space $\mathcal{X}_i'(p')$ with same probability β. Then,

$$\arg\min_{B \in \mathcal{B}} \text{Vol}(B|\beta) = \mathbf{B}$$

Proof: For a multivariate normal distribution input, to rotate the space in the direction of its p components is equivalent to solve the eigenvalue equation of covariance matrix Σ. The rotation will produce a diagonal matrix D with eigenvalues as its diagonal elements. This new matrix D is the covariance matrix of the components, which implies that they are all independent of each other and follow a normal distribution. So the problem is reduced to an instantiation of Theorem 2. Thus the result holds as before. □

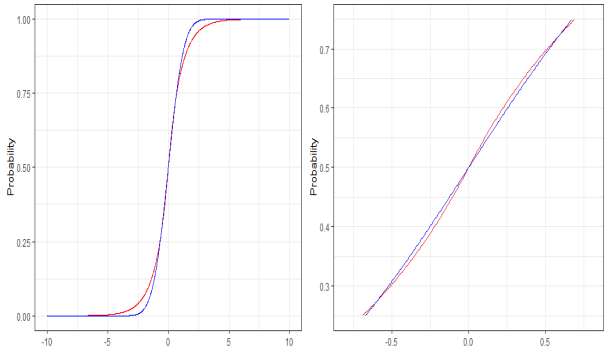
The multivariate normal distribution and its properties suggest that a more general result is attainable. It seems clear that when we have a symmetric multivariate distribution which is unimodal, and all its marginals belong to the same family of distributions, differing maybe only on the particular values of its parameters, a similar result to Theorem 3 can be obtained. For instance, Theorem 4 below shows that a similar result is obtained for a symmetric multivariate Laplace distribution. The Laplace distribution is interesting in that even in the presence of 0 correlation of the marginals, which are univariate Laplace distributions themselves, they are not independent [22, pp. 229-245]. This fact highlights an important property about our results: they do not depend on the independence of the marginals but on their zero correlation, since we only need the variables being orthogonal.

Theorem 4. Let X be a p-random vector with symmetric multivariate Laplace distribution. Assume that its components $X_1, \ldots, X_p$, being Laplace distributed, have parameters $\mathcal{L}(0, b_1), \ldots, \mathcal{L}(0, b_p)$, respectively. Assume also that $\text{corr}(X_i, X_j) = 0$ for all $i \neq j$, and $i, j \in \{1, \ldots, p\}$. Apply fastPRIM on each possible projection space $\mathcal{X}_i'(p')$ with same probability β. Then,

$$\arg\min_{B \in \mathcal{B}} \text{Vol}(B|\beta) = \mathbf{B}.$$

Proof: The proof process is the analogous to the one in Theorem 2, except that we apply it now to the Laplace distribution context. The goal is again to prove that smaller variance components will give smaller Lebesgue measures under the same probability measure $\beta^{1/p'}$. The Lebesgue measure of the j-component can be calculated by $L_j = [Q(0, b_j; 0.5 + \beta^{1/p'}) - Q(0, b_j; 0.5 - \beta^{1/p'})]$, where $Q(\mu, b; p) = \mu + b\ln(2p)$ for $p \leq 1/2$ and $\mu - b\ln(2 - 2p)$ for $p > 1/2$. Observe how variances influence L_j: a smaller variance $2b_j^2$ demands a smaller b_j. Since the quantile function Q increases for $p \leq 1/2$ and decreases for $p > 1/2$, so $Q(0, b_j; 0.5 + \beta^{1/p'})$ will decrease and $Q(0, b_j; 0.5 - \beta^{1/p'})$ will increase, which makes L_j decrease as a whole. □

Remark 1. Theorems 3 and 4 suggest that our theorem can be further generalized to a broad range of multivariate distributions. In fact, the MNIST example considered below also points in the same direction, since the distributions are somewhat deviating from normality. However, it is important to notice that distributions pertaining to different families cannot in general be mixed and the previous results maintained. Consider, for instance, X_1 having normal distribution $\mathcal{N}(0, 1)$ and X_2 having Laplace distribution $\mathcal{L}(0, 1)$ (so its variance is 2), and the two variables are independent of each other. If these two represent the marginals of some multivariate distribution and we attempt to apply our result, we will obtain that L_1 is going to be larger than L_2 when the probability measure is chosen to be small. Fig. 2 below illustrates clearly this point. However, it is also clear from Fig. 2 that there is a region $R \subset [0, 1]$ (not necessarily contiguous) such that for $\beta \in R$ it is better to select the normal distribution than the Laplace one.



(a) Distribution function (b) Zoom of distribution function

Figure 2: Counterexample for a mixed multivariate distribution. The blue line is normal $\mathcal{N}(0,1)$ and the red line is Laplace $\mathcal{L}(0,1)$.

Remark 2. Theorems 2, 3, and 4, together with Proposition 1, show that the size of the box $\mathbf{B}$ maximizes the Shannon information among all the regions of probability β in all projections $\mathcal{X}_i(p')$.

4 Simulation

We generated a dataset with 300 observations and 100 dimensions, which follows the multivariate normal distribution $\mathcal{N}_p(0,\Sigma)$. The covariance matrix Σ is assumed to have its first two diagonal elements as 1, and $\text{cov}(x_1, x_2) = 0.7$; the last two diagonal elements are 12 with $\text{cov}(x_{99}, x_{100}) = 8$; the remaining diagonal elements are 6 and all other places are 0. A pure PRIM, no rotation nor reduction, with ten iterations of covering, was then applied for contrast purposes. The result is shown in Fig. 3 for the first two dimensions. We next standardized the dataset to use the correlation matrix and apply the rotation in the direction of the eigenvectors. Once the rotation was performed, the two new variables with larger variances are the principal components, and the two variables with smaller variances are the pettiest components. The next step was to run both PRIM and fastPRIM on the two principal components and the two pettiest components, both with $\beta = 0.1$. The results are shown in Fig. 4. Each box results from a whole peeling period and the order of the boxes after ten stages of covering is shown by grading color. The fastPRIM algorithm exhibited nested rectangular boxes in contrast with the PRIM's messy ones. There is a striking difference in the result of both algorithms (PRIM and fastPRIM) between principal and pettiest components: the final boxes obtained by fastPRIM are almost five times smaller than those of its competitor.

Table 1 shows the quantitative results in terms of the empirical density over volume of the box. Looking at it by columns, we can compare the different methods under the same covering step. First, the classical PRIM results in too small densities, thus rendering it completely useless. This in the end comes as no surprise considering the curse of dimensionality. In this sense, Table 1 also reveals the benefits of dimension reduction, since any of the other options will do much better. That is, when using principal

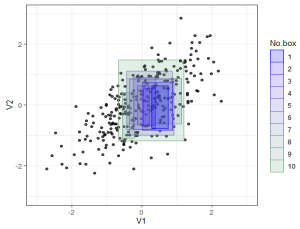
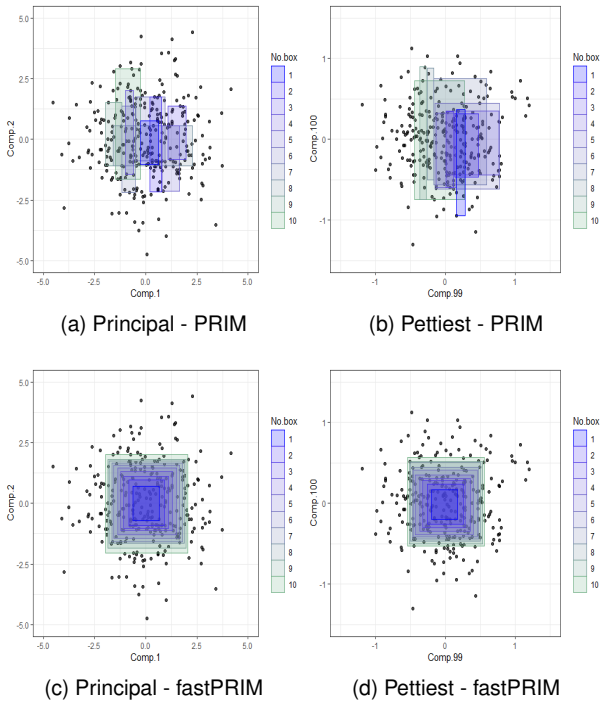


Figure 3: Region obtained by PRIM.



(a) Principal - PRIM (b) Pettiest - PRIM

(c) Principal - fastPRIM (d) Pettiest - fastPRIM

Figure 4: Simulation results by method.

or pettiest components analysis to reduce dimensionality, from 100 dimensions to 2 in our example, we get much better densities. In fact, fastPRIM improves the densities even more. However, the optimal improvement comes when the pettiest components are used; in this case, the densities show an increase of an order of magnitude. Finally, looking at the table by rows, one can track the variations of the densities during covering procedures. There is an interesting contrast in that for PRIM the density starts from a really small value and then increases monotonically; whereas for all other cases, a final decrease is accomplished as expected, though it is not monotonic.

Table 2 shows empirical bias and variance for the different methods. The bias was calculated as the distance between the mean value and the origin. The results were obtained by averaging over 100 simulations from the same distribution used before. We can see that the variance of PRIM is too high, compared to the other mechanisms with the dimension reduction, corresponding to what we would expect. Also, PRIM with principal components and fastPRIM with principal components have variances that are close, but there is a huge reduction of bias between the two methods. The same is true for PRIM and fastPRIM both

with pettiest components. This illustrates the superiority of fastPRIM in this scenario. Finally, notice that pettiest components is producing a tenfold decrease in variance, whether we are looking at PRIM or fastPRIM; the bias also decreases. The final outcome is that fastPRIM with pettiest components, having lowest variance and bias among all the strategies, minimizes the empirical mean squared error.

5 Example: MNIST Dataset

MNIST is a famous dataset of handwritten digits widely used in machine learning [23]. Using principal components analysis with the MNIST dataset is somewhat common. In fact, a few principal components have been used to reconstruct the digits [24, pp. 433-445]. With this idea in mind, we suggest that the platonic pattern of each digit is close to the mode. Therefore, here we use pettiest components and show that for PRIM and fastPRIM applied to MNIST, the pettiest components give a higher active information than the principal components. Thus, in words, the machine can use principal components in order to *read* a digit, learning all its variability (see [25, pp. 536-539], though with a different hand-written digits dataset). But if the machine is going to learn to *write* it, it is better to go for the mode in pettiest components. Thus it is better to work with pettiest components in order to generate the actual image.

Since our goal is the identification of modal patterns for digits, only the training dataset is used. The data consist of grey levels, from white to black, of 28×28 pixels for 60,000 observations. This results in a $60,000 \times 784$ matrix. All images are centered by the center of mass of the pixels. We first split the big dataset by digit to get 10 smaller datasets with size near 6000×784. Notice that the graphs are comprised mostly of white pixels, corresponding to zero value. If these white pixels are not removed, the modes will obviously be biased towards those values. Therefore these zero points make mode hunting strategies unsuitable. We need to find a threshold to make sure that most observations will be colored on those pixels. The threshold will cause a reduction in dimensionality, which should have different degrees corresponding to different numbers. The relative ranking of these reductions should be stable when the threshold changes. So we measure this reduction by the percentage of pixels we choose to keep and rank it by number from high to low. The ranking plot by threshold is shown in Fig. 5, from which it is observed that the ranking is relatively stable when the threshold is lower than 60%.

In this fashion, we obtain 10 datasets corresponding to each digit, with about 6000 observations and dimensionality smaller than 784. In order to apply PRIM and fastPRIM with principal and pettiest components analysis on each of the 10 datasets, we need first to determine for each digit the right probability β of the region that will contain the mode. We achieve this by minimizing the mean squared error through a 10-fold cross-validation, digit by digit for fastPRIM with pettiest components (Fig. 6). Then, in oder to be able to make comparisons between strategies, we use the same β-optimized value for fastPRIM with principal components, as well as for PRIM with pettiest and principal components. When the size of β permits it, we have several iterations of the covering process for PRIM, making β' the probability of

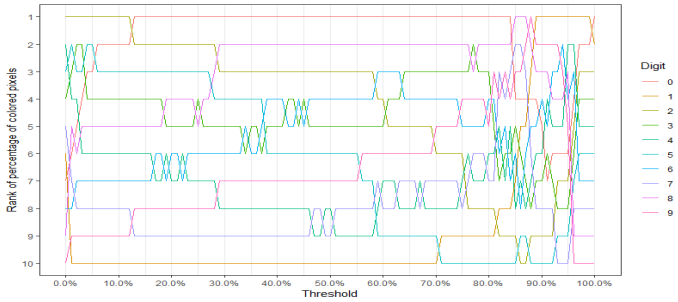


Figure 5: Ranking changes by threshold

the box in each iteration of the covering process, to obtain a final region of probability $\beta = 1 - (1 - \beta')^t$ after t iterations (Table 3). Thus, for instance, the digits 0 and 1 have just a single stage of covering, whereas the digits 8 and 9 have 7 and 4, respectively. The results are shown in Figs. 7 and 8, where the superiority of pettiest over principal components in detecting the modes is clearly observable.

To see this superiority formally, we calculate the active information (2) for each of the 10 digits considering the 4 different methods (PRIM with principal components, PRIM with pettiest components, fastPRIM with principal components, and fastPRIM with pettiest components), setting the base of the logarithm at 2. After, we rank the digits from highest to lowest in terms of the active information of the final region obtained. The results are shown in Table 4 (we recommend to look simultaneously at Fig. 9, a sample of 30 observations from the training set, to better understand this discussion).

According to (4), the β-mode is the region of the space with probability β having the highest active information. It means that this β-mode corresponds to the region of the space with minimal hyper-volume having probability β. Taking also into account that active information was first introduced to measure how much information a programmer is adding to an algorithm in order for such algorithm to reach a target T with respect to a blind search [14], [15], we can offer an interpretation of the findings in Table 4.

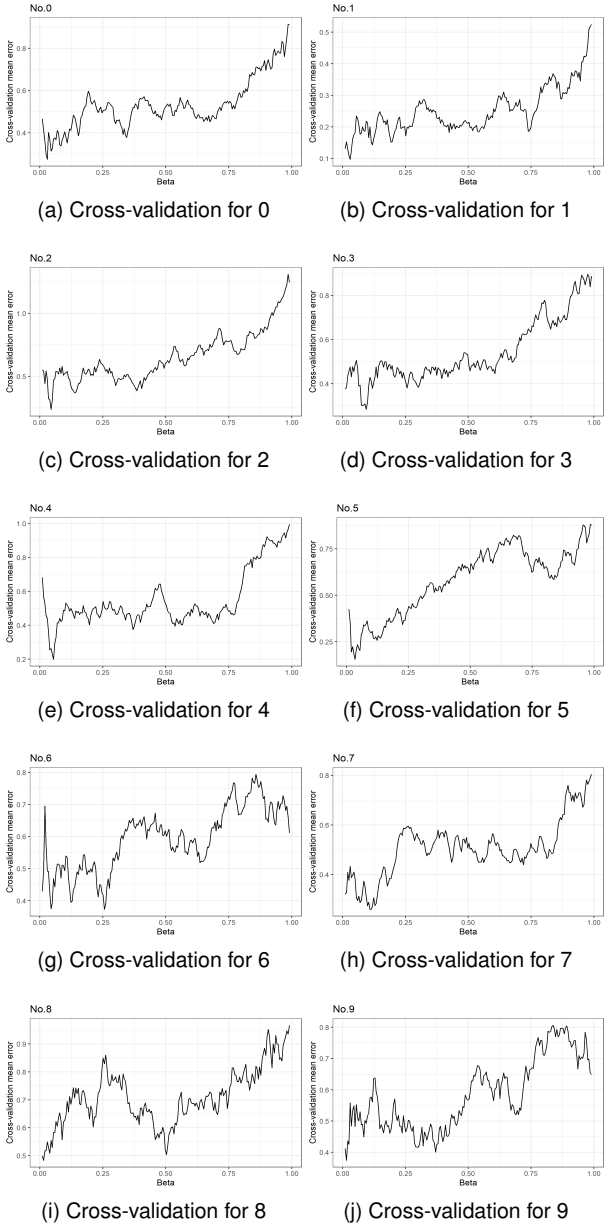
The target T in the particular context of our example is the handwritten pattern of each digit, say the Platonic idealization of Arabic numerals. Since such pattern for 1 is the easiest to follow (a vertical segment), it is not surprising that the active information of the mode representing 1 is about 8. In fact, there is a difference of almost two bits of information in the ranking of the first two digits. Accordingly, it is reasonable to expect the digit 0 to rank high, since its very well defined geometric representation makes for an easy pattern to follow (a circumference); its active information is close to 6. Somewhat surprisingly the digit 5 sits second, between the digits 1 and 0. Unsurprisingly, all the last five digits (3, 4, 6, 9, 8) have active information below 5, which is easily explainable in terms of the more complex shapes they have, since all, except the digit 4, involve some circular shape plus some additional pattern. As for 4, its ranking is possibly explained by the two ways there is to represent it. Notice also that the digits 6, 9, and 8, having obtained the highest estimation of β through the cross-validation, rank in the last positions. The digit 8 ranks

Table 1: Density of boxes per volume by different method

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PRIM	1.07e-73	1.57e-73	1.76e-73	2.14e-73	2.59e-73	3.04e-73	3.45e-73	3.87e-73	4.26e-73	4.65e-73
PRIM-Principal	15.7	16.7	14.2	14.7	14.3	14.2	14.0	13.9	13.8	13.0
fastPRIM-Principal	16.3	15.6	16.2	16.3	14.5	13.2	13.1	13.1	12.7	11.8
PRIM-Pettiest	182	168	187	152	155	142	130	132	132	127
fastPRIM-Pettiest	215	240	237	218	207	186	189	177	169	161

Table 2: Empirical variance and bias by method

	PRIM	PRIM-Principal	PRIM-Pettiest	fastPRIM-Principal	fastPRIM-Pettiest
Variance	104	2.28	0.204	2.23	0.149
Bias	0.310	0.145	0.1	0.00579	0.000653



(a) Cross-validation for 0 (b) Cross-validation for 1

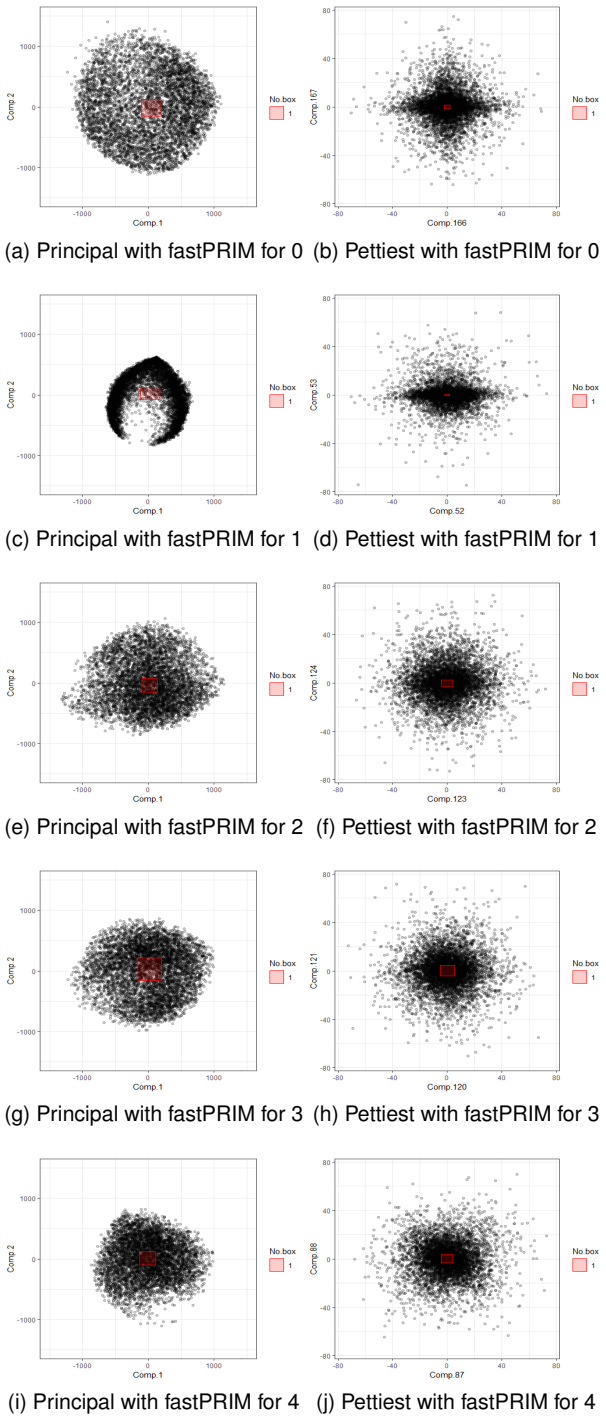
(c) Cross-validation for 2 (d) Cross-validation for 3

(e) Cross-validation for 4 (f) Cross-validation for 5

(g) Cross-validation for 6 (h) Cross-validation for 7

(i) Cross-validation for 8 (j) Cross-validation for 9

Figure 6: Cross-validation per digit



(a) Principal with fastPRIM for 0 (b) Pettiest with fastPRIM for 0

(c) Principal with fastPRIM for 1 (d) Pettiest with fastPRIM for 1

(e) Principal with fastPRIM for 2 (f) Pettiest with fastPRIM for 2

(g) Principal with fastPRIM for 3 (h) Pettiest with fastPRIM for 3

(i) Principal with fastPRIM for 4 (j) Pettiest with fastPRIM for 4

Figure 7: MNIST modeling results 0–4 (fastPRIM)

last, being the only digit with active information around 3 bits, and a difference of 1.6 bits with respect to the digit 9, the second to last. However, it is important to notice that even with 8 we see an active information bigger than 3 for the mode of the pettiest components; this says that the mode is at least 8 times more probable than the uniform

Table 3: Final region size and iterations per digit

	0	1	2	3	4	5	6	7	8	9
β	2.96%	3.21%	4.45%	9.37%	5.43%	3.46%	25.6%	11.3%	50.2%	36.9%
iterations	1	1	1	1	1	1	3	1	7	4

Table 4: Active information by method

Num.	PRIM-Principal	Num.	fastPRIM-Principal	Num.	PRIM-Pettiest	Num.	fastPRIM-Pettiest
2	1.90	2	1.83	1	8.08	1	7.96
4	1.81	4	1.79	5	6.30	5	6.15
1	1.79	7	1.36	0	6.26	0	5.97
7	1.58	8	1.32	7	5.71	7	5.71
5	1.47	3	1.24	2	5.44	2	5.42
3	1.36	9	1.07	3	4.96	3	4.94
8	1.31	5	0.89	4	4.69	4	4.70
0	1.29	6	0.81	6	4.54	6	4.54
9	1.24	0	0.60	9	4.06	9	4.13
6	1.05	1	-1.21	8	3.39	8	3.36

probability of the same box. Thus, even with 8 a very well defined pattern is followed; more so with all the other digits.

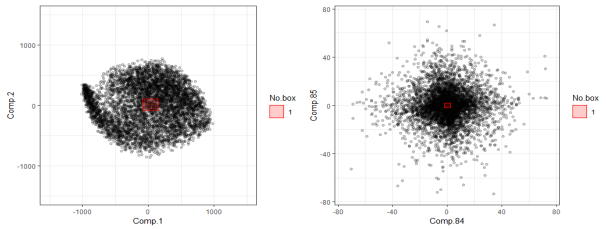
Observe that the ordering of the digits in Table 4 is almost inverted when we see them first in the principal components and then in the pettiest components (both with PRIM and fastPRIM). For instance, 1 ranks first with fastPRIM pettiest components but last with fastPRIM principal components, obtaining a negative active information. Also, 4 ranks in the second half with fastPRIM pettiest components but second with fastPRIM principal. Nonetheless, active information will clearly show that it will be mistaken to look for the β-mode of the projection in the principal components instead of the pettiest components. To see this, notice from Table 4 that we obtain an important active information gain when jumping from principal to pettiest components. Since β is fixed for each digit, and the active information of the final region R can be written as

$$\begin{aligned}\mathbf{I}_+(R) &= \log\beta - \log\mathbf{U}(R)\\ &= \log\beta + \log(\mathrm{Vol}(S)) - \log(\mathrm{Vol}(R)),\end{aligned}$$

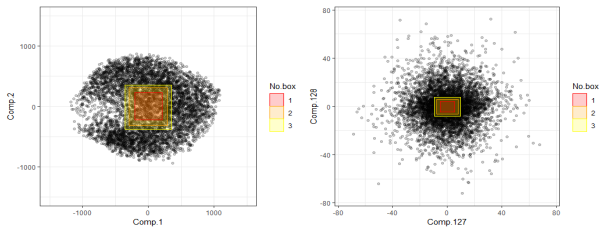
where S is the sample space, the only term that is changing in determining the active information between PRIM and fastPRIM is the last term at the RHS of the previous equation, $\log(\mathrm{Vol}(R))$. Thus, for instance, for the digit 1 evaluated with fastPRIM the final region obtained with pettiest components is more than 2^9 times smaller than the region found with fastPRIM principal components. Even for the digit 8, whose active information difference between fastPRIM pettiest and fastPRIM principal is minimal, we obtain that the final region with pettiest components is $\sim 2^4$ smaller than the region with principal components. The volume reduction is extremely significant. In fact, observe that many of the distributions around principal components look uniform (see, e.g., 7(a) and (g) for the digits 1 and 3 with principal components, with the notorious exception of the digit 1), whereas the distribution of digits with pettiest components have in general more structure that gives them starry shapes.

Notice also the interesting fact that when looking at the pettiest components, PRIM does better than fastPRIM with the digits 0, 1, and 5. This can be puzzling at the beginning. Comparing fastPRIM and PRIM with principal components for the digit 1, the most extreme case, is enlightening here. Notice from Figs. 7(c) and 8(c), that the projected distribution in the dimension of the two principal components is not symmetric. Since fastPRIM was developed for symmetric unimodal distributions, it is not surprising that PRIM does a better job in this case. Nonetheless, it is important to notice that there is no violation of the results obtained in our theorems, since the conditions are not satisfied. Nonetheless, it is important to highlight that in spite of the data not being normal or Laplace, PRIM with pettiest components is doing only slightly better than fastPRIM, illustrating that even in these cases few to nothing is lost if we consider fastPRIM instead of PRIM.

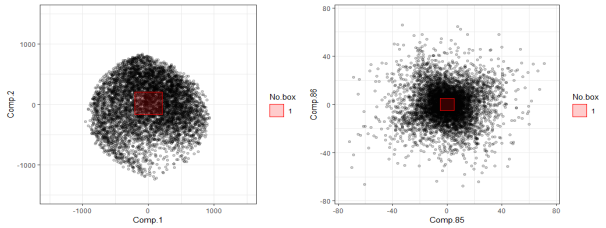
Finally, we propose to use pettiest components to reconstruct the number. Or, to be more accurate, we propose using the region with the highest active information in order to reconstruct the number. Our reasoning is simple. The region with the largest active information will be close to the platonic idealization of the digit (or at least a democratized version of such platonic idealization). To see this, we show here in Fig. 11 how do the digits for 1, 5, 0, and 8 look between fastPRIM with pettiest and principal components. Notice that the higher the difference in active information between the two procedures, the more obvious become to use of this strategy. For instance, for digit 1 the difference in active information between fastPRIM pettiest and principal components is above 9 bits, and it is very clear that the digits written with pettiest components are more homogeneous, the same is true for the digits 0 and 5, whose difference in active information between fastPRIM pettiest and principal is over 5 bits. Following this trend we end with 8, whose difference of active information between the two procedures is around 2 bits and visually the difference is not that obvious. Therefore, we claim that it is better to use the region with the highest active information to reconstruct the image, in this case any of the procedures considered with pettiest components. Fig. 10 illustrates this point comparing reconstruction between fastPRIM with pettiest and fastPRIM with principal: 0 and 1 seem bolder with pettiest; 2, 7, and even 9 look better finished with pettiest; 3 has a more defined round form in the lower part with pettiest, similar to what was discovered in [25, pp. 536-539] with principal components, but we show here that it is even better with pettiest; 5 with principal looks more like and 'S', while 5 with pettiest looks better defined; 6 seems a slightly better with principal; and 8 does not seem to show too many visible differences, but when the image is zoomed in the center looks neater with pettiest.



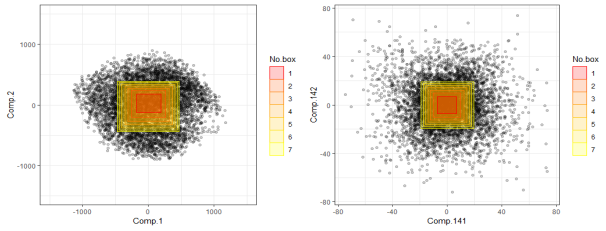
(k) Principal with fastPRIM for 5 (l) Pettiest with fastPRIM for 5



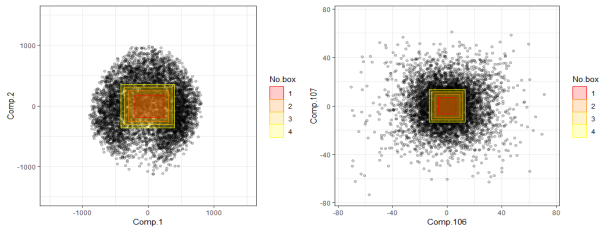
(m) Principal with fastPRIM for 6 (n) Pettiest with fastPRIM for 6



(o) Principal with fastPRIM for 7 (p) Pettiest with fastPRIM for 7

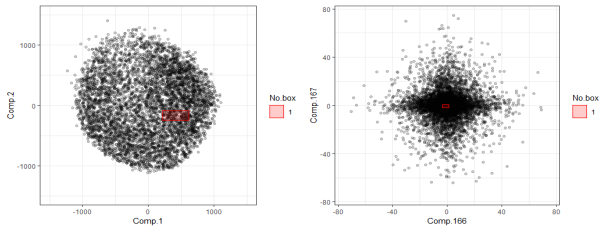


(q) Principal with fastPRIM for 8 (r) Pettiest with fastPRIM for 8

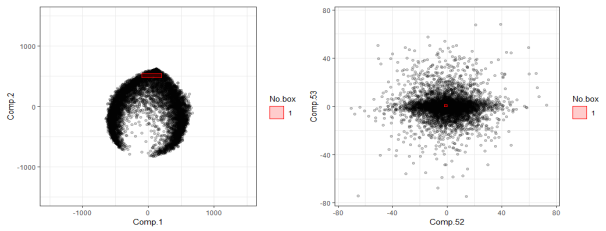


(s) Principal with fastPRIM for 9 (t) Pettiest with fastPRIM for 9

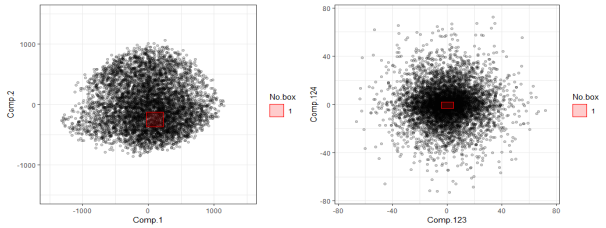
Figure 7: MNIST modeling results 5–9 (fastPRIM)



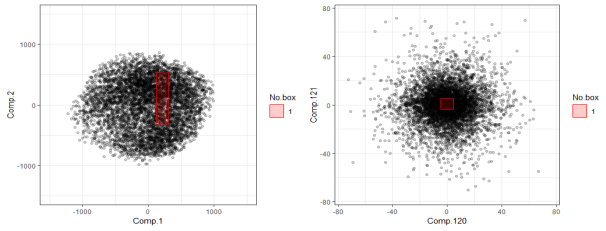
(a) Principal with PRIM for 0 (b) Pettiest with PRIM for 0



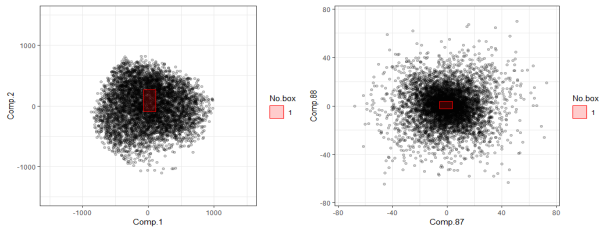
(c) Principal with PRIM for 1 (d) Pettiest with PRIM for 1



(e) Principal with PRIM for 2 (f) Pettiest with PRIM for 2



(g) Principal with PRIM for 3 (h) Pettiest with PRIM for 3



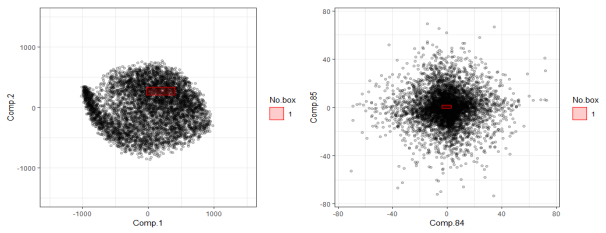
(i) Principal with PRIM for 4 (j) Pettiest with PRIM for 4

Figure 8: MNIST modeling results 0–4 (PRIM)

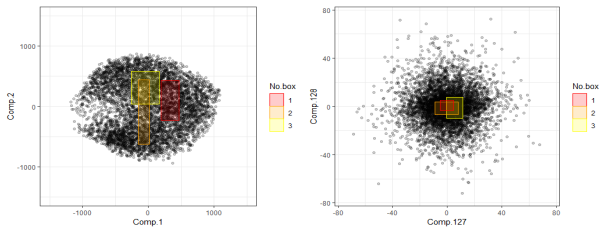
6 Summary

Given the mean's lack of robustness even in low dimensions, and the prevalence of big data and large-dimensions analyses, the importance of mode-based statistics and learning is becoming more relevant [26]. Consequently, new theory and methods are required in order to better detect modes, but this task is difficult and elusive since kernel functions become ineffective in even not-that-high dimensions. The importance of fastPRIM in mode hunting is dictated by the central limit theorem. The setback of this approach is the curse of dimensionality, as Vapnik explains in the Introduction of his book [27, pp. 4–6]. Therefore, even though there is some optimality that is reached just by rotating the information in the direction of the eigenvalues, if mode detection methods are going to be useful they will need to reduce dimensionality.

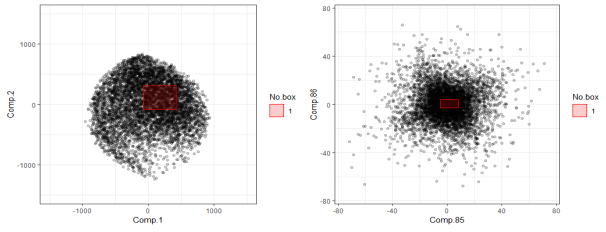
In this sense, even though it is well known that pettiest components can sometimes explain better a response than principal components, the latter have been treated as the ideal tool whereas the former have been considered isolated



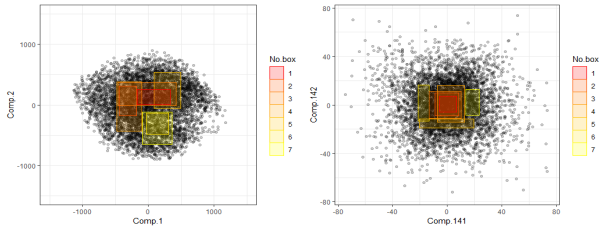
(k) Principal with PRIM for 5 (l) Pettiest with PRIM for 5



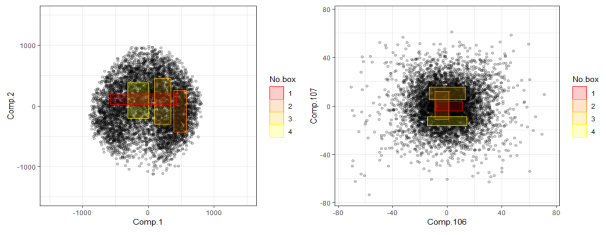
(m) Principal with PRIM for 6 (n) Pettiest with PRIM for 6



(o) Principal with PRIM for 7 (p) Pettiest with PRIM for 7



(q) Principal with PRIM for 8 (r) Pettiest with PRIM for 8



(s) Principal with PRIM for 9 (t) Pettiest with PRIM for 9

Figure 8: MNIST modeling results 5–9 (PRIM)

counter-examples or anomalies to be avoided. However, in this article we showed that when projecting the multivariate data in the direction of a few eigenvalues, pettiest components can be systematically used in order to find the best β-modes, provided that the data is distributed normal or Laplace. In this sense, the fact that it had not been noted before is surprising, given the centrality of the normal distribution in statistics.

This finding goes against the general notion that principal components are more informative, since, as shown in the

Figure 9: Handwritten digits samples

(a) Reconstruction with principal

(b) Reconstruction with pettiest

Figure 10: Reconstructed digits with fastPRIM.

Introduction, the box with the smallest size also maximizes the active information relative to the uniform distribution. In fact, not only the theorems, but the example with the MNIST dataset illustrates that principal components can totally twist the importance of the components, since active information shows that the true order is the one given by pettiest components; and this is so even when there are strong departures from normality, using pettiest over principal components produces significant gains.

Some questions remain open. For instance, we might ask how much can Theorems 3 and 4 be extended so that they become particular cases of a more general result in which we are considering unimodal symmetric multivariate distributions formed by marginals corresponding to the same family of random variables (in our case we considered all marginals being normal or Laplace). We might also ask to what extent can the marginals be from families of distribution (as in the counterexample), while our result is maintained.

In spite of these open questions, as the real data example shows, using pettiest components instead of principal components in order to determine the best β-mode in a projected space is, in general, a wise idea.

References

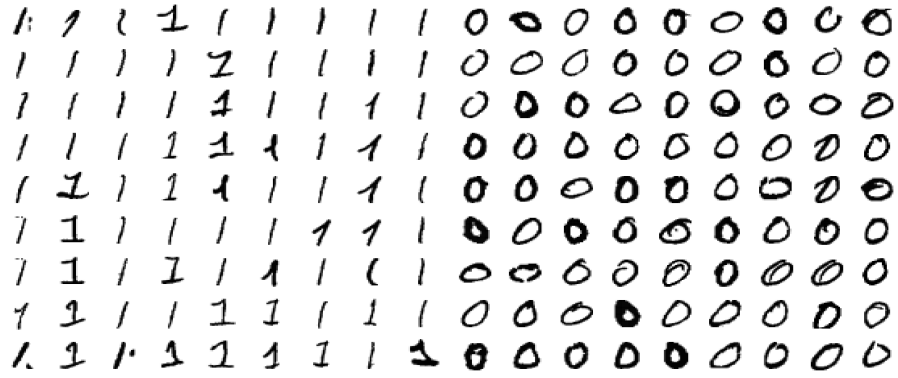
[1] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[2] A. Artemiou and B. Li, "On principal components and regression: A statistical explanation of a natural phenomenon," *Stat. Sinica*, vol. 19, pp. 1557–1565, 2009. [Online]. Available: www.jstor.org/stable/24308917

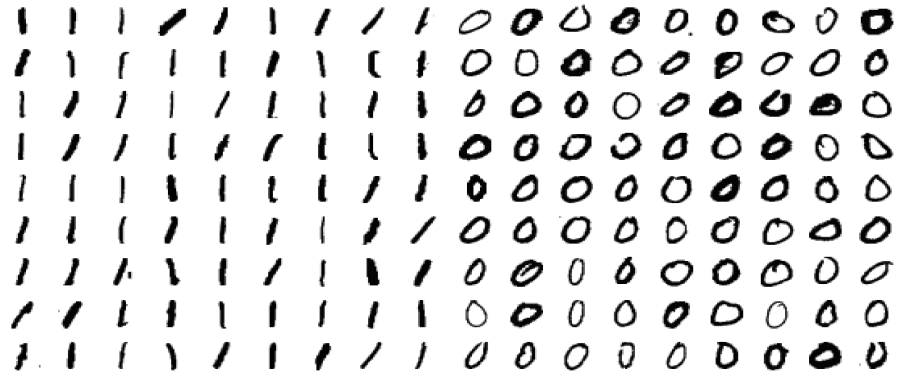
[3] I. T. Joliffe, "A Note on the Use of Principal Components in Regression," *J. Roy. Stat. Soc. C*, vol. 31, no. 3, pp. 300–303, 1982. [Online]. Available: https://doi.org/10.2307/2348005

[4] A. S. Hadi and R. F. Ling, "Some Cautionary Notes on the Use of Principal Components Regression," *Am. Stat.*, vol. 52, no. 1, pp. 15–19, 1998. [Online]. Available: https://doi.org/10.1080/00031305.1998.10480530

[5] D. R. Cox, "Notes on some aspects of regression analysis," *J. Roy. Stat. Soc. A*, vol. 131, p. 265/279, 1968. [Online]. Available: https://doi.org/10.2307/2343523



(a) Grid of 1 with principal (b) Grid of 0 with principal



(c) Grid of 1 with pettiest (d) Grid of 0 with pettiest

(e) Grid of 5 with principal (f) Grid of 8 with principal

(g) Grid of 5 with pettiest (h) Grid of 8 with pettiest

Figure 11: Equidistant points inside the β-region with fastPRIM

[6] K. Sando and H. Hino, "Modal principal component analysis," *Neural Computation*, vol. 32, no. 10, pp. 1901–1935, 2020. [Online]. Available: https://doi.org/10.1162/neco_a_01308

[7] D. A. Díaz-Pachón, J.-E. Dazard, and J. S. Rao, "Unsupervised Bump Hunting Using Principal Components," in *Big and Complex Data Analysis: Methodologies and Applications*, S. E. Ahmed, Ed. Springer International Publishing, 2017, pp. 325–345. [Online]. Available: https://doi.org/10.1007/978-3-319-41573-4_16

[8] K. Fukunaga and L. D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Transactions on Information Theory*, vol. 21, pp. 32–40, 1975. [Online]. Available: https://doi.org/10.1109/TIT.1975.1055330

[9] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962. [Online]. Available: https://doi.org/10.1214/aoms/1177704472

[10] P. de Valpine, "Monte Carlo State-Space Likelihoods by Weighted Posterior Kernel Density Estimation," *J. Amer. Statist. Assoc.*, vol. 99, no. 466, pp. 523–536, 2004. [Online]. Available: https://doi.org/10.1198/016214504000000476

[11] M. Cule, R. Samworth, and M. Stewart, "Maximum likelihood estimation of a multi-dimensional log-concave density," *J. Roy. Stat. Soc. B*, vol. 72, pp. 545–600, 2010. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2010.00753.x

[12] P. S. Ruzankin and A. V. Logashov, "A fast mode estimator in multidimensional space," *Stat. & Probab. Letters*, vol. 158, p. 108670, 2020. [Online]. Available: https://doi.org/10.1016/j.spl.2019.108670

[13] D. A. Díaz-Pachón, J. P. Sáenz, and J. S. Rao, "Hypothesis testing with active information," *Stat. & Probab. Letters*, vol. 161, p. 108742, 2020. [Online]. Available: https://doi.org/10.1016/j.spl.2020.108742

[14] W. A. Dembski and R. J. Marks II, "Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search," in *Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX*, October 2009, pp. 2647–2652. [Online]. Available: https://doi.org/10.1109/ICSMC.2009.5346119

[15] ——, "Conservation of Information in Search: Measuring the Cost of Success," *IEEE Transactions on Systems, Man and Cybernetics A, Systems & Humans*, vol. 5, no. 5, pp. 1051–1061, September 2009. [Online]. Available: https://doi.org/10.1109/TSMCA.2009.2025027

[16] D. A. Díaz-Pachón and R. J. Marks II, "Generalized active information: Extensions to unbounded domains," *BIO-Complexity*, vol. 2020, no. 3, pp. 1–6, 2020. [Online]. Available: https://doi.org/10.5048/BIO-C.2020.3

[17] D. A. Díaz-Pachón, J. P. Sáenz, J. S. Rao, and J.-E. Dazard, "Mode hunting through active information," *Applied Stochastic Models in Business and Industry*, vol. 35, no. 2, pp. 376–393, 2019. [Online]. Available: https://doi.org/10.1002/asmb.2430

[18] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Stat. Comput.*, vol. 9, pp. 123–143, 1999. [Online]. Available: https://doi.org/10.1023/A:1008894516817

[19] W. Polonik and Z. Wang, "PRIM analysis," *J. Multivar. Anal.*, vol. 101, no. 3, pp. 525–540, 2010. [Online]. Available: https://doi.org/10.1016/j.jmva.2009.08.010

[20] J.-E. Dazard and J. S. Rao, "Local Sparse Bump Hunting," *J. Comput. Graph. Stat.*, vol. 19, no. 4, pp. 900–929, 2010. [Online]. Available: https://doi.org/10.1198/jcgs.2010.09029

[21] J.-E. Dazard, J. S. Rao, and S. Markowitz, "Local Sparse Bump Hunting reveals molecular heterogeneity of colon tumors," *Stat. Med.*, vol. 31, no. 11-12, pp. 1203–1220, 2012. [Online]. Available: https://doi.org/10.1002/sim.4389

[22] S. Kotz, T. J. Kozubowski, and K. Podgorski, *The Laplace Distribution and Generalizations*. Birkhauser, 2001.

[23] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[24] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, 2016.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science, 2009.

[26] J. E. Chacón, "The Modal Age of Statistics," *International Statistical Review*, vol. 88, pp. 122–141, 2020. [Online]. Available: https://doi.org/10.1111/insr.12340

[27] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

Tianhao Liu received his BS in Physics from Nankai University, China, in 2019, and a MS in Biostatistics from University of Miami, in 2020. He is currently working towards a PhD in Biostatistics at the University of Miami. His research interests are mainly in statistical techniques for pattern recognition and their applications to medicine.

Daniel Andrés Díaz-Pachón received his B.S. in Mathematical Statistics at *Universidad Nacional de Colombia*, Colombia (2005); and his PhD in probability theory at *Universidade de São Paulo*, Brazil (2009).

In 2011he moved to the University of Miami, Florida, where he was first a Postdoctoral Associate in Biostatistics (2011–2015), and then became Research Assistant Professor. His research is focused on the intersection of probability theory, statistics, machine learning, and information theory.

J. Sunil Rao PhD is Professor and Director of the Division of Biostatistics in the Department of Public Health Sciences, University of Miami Miller School of Medicine. His research interests include mixed model prediction and selection, Bayesian model selection, small area estimation, machine learning and applied biostatistics.

Jean-Eudes Dazard PhD received his PhD in Bioinformatics in 2000 from the University of Montpellier, France, following two master's degrees: in computer science in 1992 from ESIM School of Engineering, France and in Statistics in 2009 from Case Western Reserve University, USA. His research is centered on Computational/Statistical Biology. Recent focus has been in: Bump Hunting, Regularization and Variance Stabilization, Variable Selection; and Causal Regulatory Network Analysis. His interest is also in Statistical Computing and Software Development: He has authored several R packages available in CRAN and GitHub.